International Journal of Electrical and Computer Engineering (IJECE)
Vol. 13, No. 4, August 2023, pp. 4256~4267
ISSN: 2088-8708, DOI: 10.11591/ijece.v13i4.pp4256-4267

An improved extrinsic monolingual plagiarism detection approach of the Bengali text

Adil Ahnaf, Hossain Mohammad Mahmudul Hasan, Nabila Sabrin Sworna, Nahid Hossain
Department of Computer Science and Engineering, United International University, Dhaka, Bangladesh

Article Info

Article history:

Received Apr 9, 2022
Revised Oct 30, 2022
Accepted Nov 6, 2022

Keywords:

Bengali language
Corpus
Detection
Plagiarism
Similarity

ABSTRACT

Plagiarism is an act of literature fraud, which is presenting others' work or ideas without giving credit to the original work. All published and unpublished written documents are under the cover of this definition. Plagiarism, which increased significantly over the last few years, is a concerning issue for students, academicians, and professionals. Due to this, there are several plagiarism detection tools or software available to detect plagiarism in different languages. Unfortunately, negligible work has been done and no plagiarism detection software available in the Bengali language where Bengali is one of the most spoken languages in the world. In this paper, we have proposed a plagiarism detection tool for the Bengali language that mainly focuses on the educational and newspaper domain. We have collected 82 textbooks from the National Curriculum of Textbooks (NCTB), Bangladesh, scrapped all articles from 12 reputed newspapers and compiled our corpus with more than 10 million sentences. The proposed method on Bengali text corpus shows an accuracy rate of 97.31%

Corresponding Author:

Nahid Hossain
Department of Computer Science and Engineering, United International University
United City, Badda, Dhaka 1212, Bangladesh
Email: nahid@cse.uiu.ac.bd

1. INTRODUCTION

Data overload in media like digital documents is expanding as information technology advances. In today's environment, digital papers must have uniqueness and proper ownership to be original. However, stealing others' work from online text documents is now so common and it has recently become a major concern for both writers and publishers. Plagiarists try to pass off others' work, ideas, or words as their own. Plagiarism can take several forms, such as copying huge portions of text or paraphrasing and replacing original words with similar or equivalent terminology. Examining a document's similarity to original documents can detect plagiarism. Plagiarism is not a violent crime, though it is unethical. Plagiarism software like Turnitin is widely used in universities. Others include Viper, eTBLAST, EVE2, Cross-Check, and iThenticate. While commercial and non-commercial plagiarism detection software is available in English, it cannot identify plagiarism in Bengali articles which is a widely spoken language in the world. We investigated and discovered that all the plagiarism detection tools available do not support the Bengali language. This motivates us to work on a plagiarism detection approach that works solely for Bengali text.

Text-based plagiarism detection systems (PDS) are mainly two types i.e., intrinsic detection and extrinsic detection. An extrinsic PDS has a reference corpus to compare with a suspicious document. However, unlike extrinsic PDS, intrinsic PDS detects plagiarism single-handedly without any corpus only

based on the writing styles of the author, structural distributions, and vocabulary richness [1]. This study mainly focused on the extrinsic plagiarism detection approach. The system creates a vector for each sentence and each vector is created by several points. Each point of every vector represents a word of their corresponding sentence. By applying the cosine similarity algorithm, the system evaluates the similarity between two vectors [2], [3]. If most of the terms are the same when comparing two sentences, then they are considered identical [4], but similar words might not necessarily signify the same thing. Thus word linguistic information is vital [5]. Paraphrasing is another way of plagiarism where a person replaces some lexemes in a sentence keeping the same meaning to avoid plagiarism. The proposed method handles paraphrasing by providing a lexicon of lexeme synonyms and creating multiple input sentences by fetching different combinations of lexeme synonyms from the lexicon. In this way, the proposed system measures the similarity of all possible combinations of a sentence with our corpus. The proposed work has two major contributions. Firstly, a highly accurate plagiarism detection tool for the Bengali language has been developed. Secondly, a sufficiently large Bengali monolingual corpus on textbooks and newspaper domains has been built, and finally, the design, codes, and corpora will be uploaded to a public repository for future researchers to contribute to the enrichment of the Bengali language.

Although there is no notable plagiarism detection approach available in the Bengali language to study or compare with, we have studied some relevant Bengali work and some English plagiarism detection approaches. Islam *et al.* [6] applied a stylometric method to determine authorship in Bengali texts. They used n-grams to classify authorship and tested their model using 3,125 passages written by ten Bengali authors. Their device achieved a precision of 96%. Pandit *et al.* [7] explored semantic similarity measures in the Bengali language. They used word2vec semantic similarity calculation and used English as their cross-lingual counterpart. Mridha *et al* [8]. proposed a writer identification system for Indic scripts including Bengali. The approach is based on non-trainable Gabor filters fused convolutional neural network (CNN). Hossain *et al.* [9] proposed a bidirectional conversion system between Chittagonian and standard Bangla. Using binary search, word-to-word mapping, and morphological transformation developed a bidirectional lexicon. Their system gained 95.86% and 93.89% accuracy rates with respectively Chittagonian to standard Bangla and standard Bangla to Chittagonian.

The following PDS in English and other western languages give us a better understanding of the plagiarism detection system. Oktoveri *et al.* [10] suggested a method for eliminating irrelevant documents that do not share a common subject with the query text. Their proposed method used a variety of algorithms and methods, including a winning algorithm; an Adelson-Velsky and Landis tree for document indexing; and least common subsequence and term frequency. Mahdavi *et al.* [11] suggested a model using external vector space (VSM) to detect plagiarism. The proposed method involves three key phases: preparing data, retrieving relevant records, and matching comprehensive strings. The pre-processing process involves six sub-steps: normalization of text, elimination of stop terms, stemming, substitution of synonyms, tokenization, and selection of features. Ravi *et al.* [12] proposed a fuzzy C means clustering algorithm and observed that this algorithm performed better for external plagiarism detection. They tested their system using PAN 2013 Corpus. The output of their method is compared to the output of other methods that use K mean clustering and N-gram. Paul and Jamal [13] suggested a plagiarism detection technique. Their work consisted of 5 stages from pre-processing to similarity detection. Rahmatulloh *et al*. [14] demonstrated a difference between the Nazief-Adriani and stemmer porter in measuring plagiarism. This study indicates Nazief-Adriani performs superior to the stemmer porter. Roostaee *et al.* [15] proposed a two-level matching scheme for plagiarism detection. They used a weighted multilingual word embeddings and graph-of-words representation of text to understand the relationship between words in a sentence. Hambi and Benabbou [16] proposed a comparison of semantic plagiarism detection methods. In this paper, they focus on deep learning-based techniques such as vector representation method, level treatment, similarity method, and dataset. Darmalaksana *et al.* [17] also developed a search engine technology that finds information from eleven hadith in the Indonesian language efficiently. The system used latent semantic analysis and cosine similarity algorithm for structured representation of text data and to evaluate the similarity between keyword text and hadith text data. Kaur *et al.* [18] developed a semantic based approach for detecting English monolingual plagiarism. They used PAN-PC-11 and PAN-14 datasets for training and testing purposes, respectively. Alvi *et al.* [19] proposed a paraphrase identification approach and plagiarism detection tool based on contexts and word embeddings. Son *et al.* [20] proposed a plagiarism detection approach using feature extraction techniques which is based on multi-layer long-short term memory (LSTM) networks. The system has two-phase to determine plagiarism strings between two documents. The method evaluated by PAN 2014 text alignment. Hourrane and Benlahmar [21] proposed a cross-lingual plagiarism detection approach based on knowledge graphs. Their system detects semantic textual similarities based on knowledge graph representations which are referred to as CL-GTA.

The paper has the following organizations. Section 2 demonstrates the proposed method. The experiment and result analysis are depicted in section 3. Finally, section 4 concludes with some final thoughts and highlights some future work of this study.

2. METHOD

Plagiarism detection techniques are classified into two groups based on language: monolingual (e.g., English-English) and cross-lingual (e.g. English-Spanish) [4], [22]. In this study, we propose a monolingual plagiarism detection tool for the Bengali language. We have used the cosine similarity measure algorithm to determine the similarity between different Bengali texts. We have studied different text similarity measure algorithms i.e., cosine similarity, Levenshtein edit distance, Manhattan distance, and Jaccard similarity. However, the Cosine similarity measure algorithm works faster and gives more accurate results compared to other algorithms on Bengali sentences consisting of multi-byte Bengali characters. The term frequency-inverse document frequency (TF-IDF) algorithm has been used as the vectorization algorithm.

2.1. Cosine similarity

Cosine similarity is a statistic that can be used to determine the similarity of a data object. The data objects in a dataset are treated as vectors in cosine similarity. It helps to determine the cosine angle. When the result is bound to [0, 1], cosine similarity is particularly effective [19]. The cosine similarity of the two vectors in the same orientation is 1, and the relative 90 orientation is 0. The Euclidean dot product formula can be used to find cosine similarity equations as in (1) and (2),

$$\vec{A}.\vec{B} = \left|\vec{A}\right|\left|\vec{B}\right| \cos\theta \tag{1}$$

$$Similarity = \cos\theta = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{2}$$

where A_i and B_i are components of vector A and B, respectively.

2.2. TF-IDF vectorization

As previously stated, the dot product of the vector representation of two sentences can be used to determine their cosine similarity. As a result, TF-IDF employs to represent the vector. It is a statistical tool for measuring the significance of a word in a collection of text [23]. It uses the decision-making rule as in (3),

$$TF - IDF(t,d) = TF(t,d) \times IDF(t) \tag{3}$$

where t denotes a specific word and d denotes the document's total number of documents. *TF* is used to calculate the frequency of a word in each text in the corpus. The number of times a word appears in a document refers to the percentage of the total number of words in that document [24]. TF's simplified formula is as (4),

$$TF(t,d) = t \times \frac{1}{d} \tag{4}$$

where t refers to the overall number of instances of a phrase in a document, and the total number of documents is denoted by d. The weight of unusual terms in all corpus documents are calculated using IDF. The formula is given in (5),

$$IDF(t) = \log\left\{(1+n) \times \frac{1}{1+df(t)}\right\} + 1 \tag{5}$$

where n indicates the total number of documents in the document set and $df(t)$ indicates the total number of documents that include the word. TF-IDF converts text into a meaningful numerical representation. It was chosen for this investigation because it considers the entire word count. The TF-IDF algorithm assigns great weight to documents with high TF but low TF. It works by boosting the frequency of a term in a document. Table 1 demonstrates an example with dummy articles (A1 to A3) to get better insights. On the other hand, Figure 1 demonstrates a vector graph of A1, A2, and A3.

A1='*সততা শিক্ষা এবং শিক্ষা সততা*(Honesty is education and education is honesty)'
A2='*শিক্ষাক্ষেত্রে সততা থাকা জরুরি*(Honesty is important in education)'
A3='*ব্যক্তি জীবনেও সততা থাকা জরুরি*(Honesty is also important in personal life)'

Table 1. Vectorization by using TF-IDF

Term	TF Value			IDF Value	TF-IDF Value		
	A1	A2	A3		A1	A2	A3
সততা (honesty)	2/5	1/4	1/5	0	0	0	0
শিক্ষা (education)	2/5	0/4	0/5	0.477	0.191	0	0
এবং (and)	1/5	0/4	0/5	0.477	0.095	0	0
শিক্ষাক্ষেত্রে (in education)	0/5	1/4	0/5	0.477	0	0.119	0
থাকা (remain)	0/5	1/4	1/5	0.176	0	0.044	0.044
জরুরি (important)	0/5	1/4	1/5	0.176	0	0.044	0.044
ব্যক্তি (person)	0/5	0/4	1/5	0.477	0	0	0.095
জীবনেও (life too)	0/5	0/4	1/5	0.477	0	0	0.095

Previously, we discussed how vector sentence identification distinguishes between sentences that are similar or identical. It occurs when the system turns the source text and suspicious patterns into a similarity vector using TF-IDF vectorization. The graph in Figure 1 depicts the overlapping regions; a text appears suspicious, especially if it closely resembles the source. Because the vector parameter plots between sentences A2 and A3 are so close, the two sentences seem to be related.

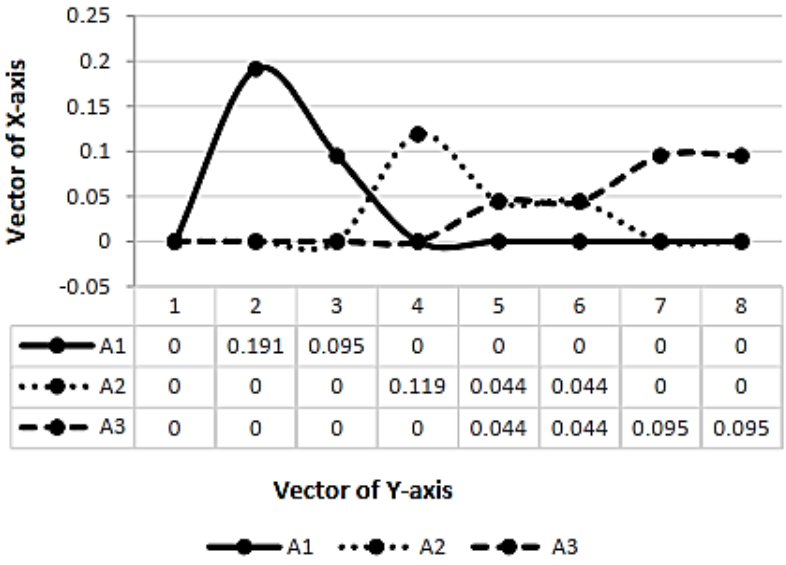

Figure 1. Vector graph of A1, A2, and A3

As mentioned earlier, this project intends to develop monolingual plagiarism detection algorithms (Bengali to Bengali). In this project, we cover two distinct domains i.e., education and news. This section covers data collection, corpus creation, and the process of plagiarism detection on Bengali texts. Figure 2 demonstrates the architecture of plagiarism detection.

2.3. Data collection

Since our plagiarism detection approach is extrinsic, data serves as the principal component. It is usually in written Unicode Bengali text format. In our work, the entire dataset construction process is divided into two parts, which are educational domain and newspaper domain.

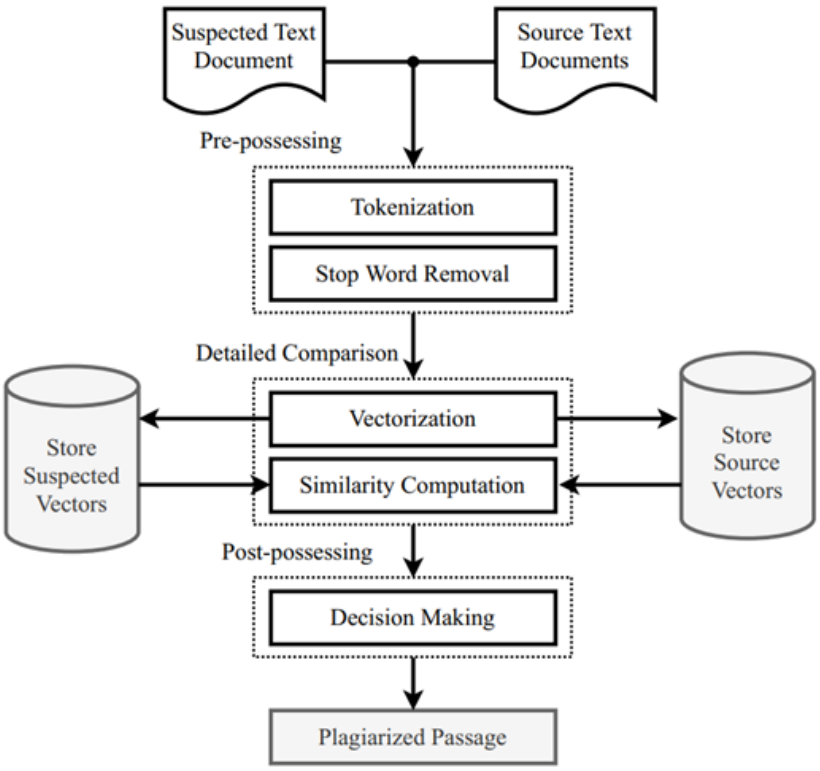

Figure 2. The architecture of plagiarism detection

2.3.1. Educational domain

An educational domain dataset has been created by collecting all the books of class I to XII except the English literature books available on the NCTB website [25]. NCTB textbooks are not available in Unicode format. It is available in PDF format. In this work, we have converted all pdf files into a writable Docs format. Furthermore, we have removed all the figures, tables, extra whitespace, and tabs from the book. Due to this, it provides data that contains plain Bengali text from the book. This is a manual process, and it took around 2 months to complete. In this way, the data for each book has been stored in a separate text file. The educational domain dataset formation process is depicted in Figure 3.

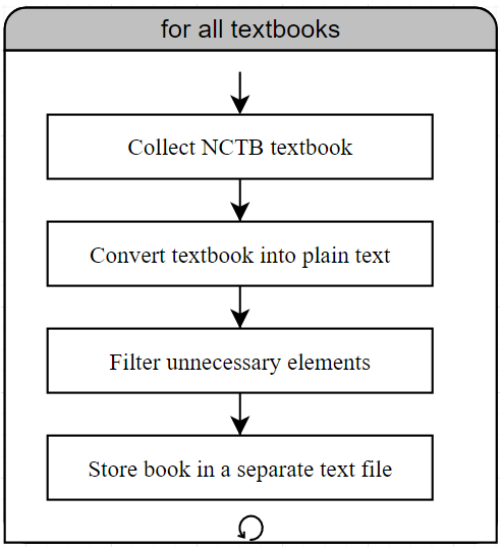

Figure 3. Educational domain data collection

2.3.2. Newspaper domain

In this study, the Scrapy framework of python has been used to build different spider bots to collect news texts from different Bengali online newspapers to create a dataset in the newspaper domain. News articles have been collected in equal proportion from 8 highly well-known and prestigious Bangladeshi online newspapers namely Prothomalo, Bdnews24, Banglanews24, Samakal, Kalerkantho, Jugantor, Inqilab, Manabzamin to make the newspaper corpus balanced and unbiased. The crawler crawls all articles of a particular newspaper and scrapes all the content from the articles. Specific Bengali text is filtered and then embedded in a text file. This process took more than 1 month. Moreover, we have taken nearly 2 million sentences from the reputed Bengali monolingual corpus NHMono01 [26] in the newspaper domain. The complete newspaper dataset formation process is shown in Figure 4.

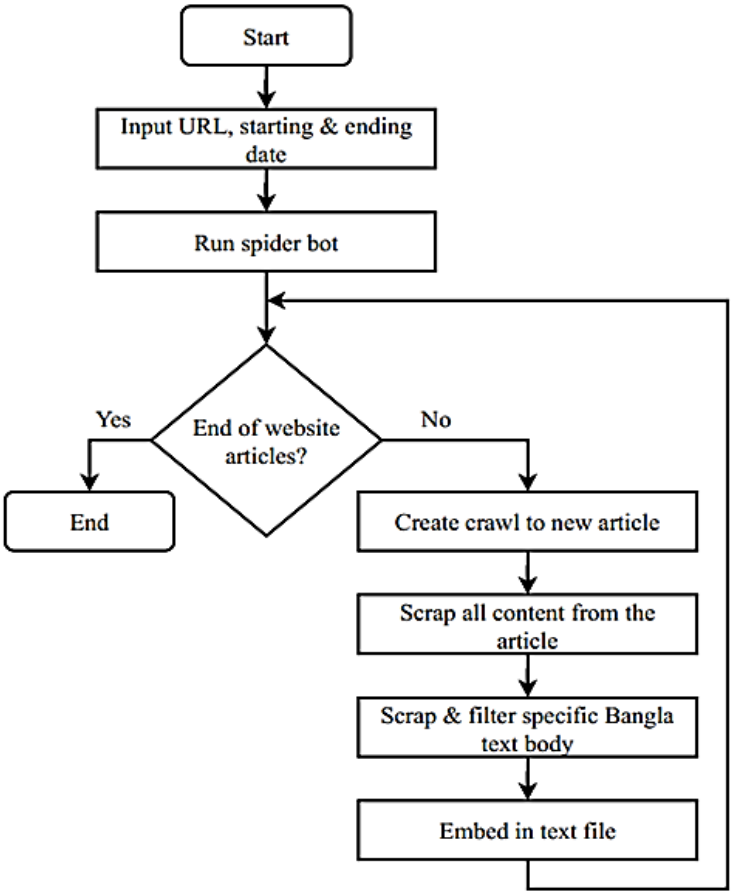

Figure 4. Newspaper domain data collection

2.3.3. Lexicon for paraphrasing

As mentioned earlier, a lexicon has been developed that contains synonyms of lexemes to create possible sentences from each input sentence to tackle paraphrasing. First, unique lexemes from both educational and newspaper domains are collected. Then, for each lexeme, all the synonyms of that lexeme are collected and stored in the lexicon. The synonyms of these Bengali lexemes are collected from Bengali to Bengali dictionary published by Bangla Academy which is funded by the government of Bangladesh [27]. Finally, the authors spent an ample amount of time proofreading the whole lexicon for any discrepancies. The final lexicon contains synonyms for 80,134 unique Bengali words. Table 2 demonstrates some samples of the lexicon.

2.4. Prepossessing and corpus creation

Before feeding data to the model, each of these data needs to be preprocessed. In order to do that, first, the tokenizer is used to tokenize the sentences into words, then the stop words are removed. Finally, similarities between the sentences are generated to compare and find out plagiarism in the article.

Table 2. Samples from synonym lexicon

Lexeme	Synonyms
মূর্খ (fool)	*অজ্ঞ* (ignorant), *অশিক্ষিত* (uneducated), *নির্বোধ* (silly), *বেকুব* (stupid)
দুঃসময় (bad days)	*অকাল* (ill-timed), *অসময়* (untimely), *কুদিন* (bad day), *দুঃসময়* (bad time)
অঙ্গীকার (commitment)	*পণ* (oath), *প্রতিশ্রুতি* (promise), *প্রতিজ্ঞা* (swear), *শপথ* (oath)
অধ্যয়ন (specialism)	*পাঠ* (lesson), *পঠন* (reading), *পড়া* (study), *লেখাপড়া* (education)

2.4.1. Tokenization of the sentences

The process of splitting each word from a sentence based on a particular delimiter i.e., punctuation, new lines, tabs, or characters is called Tokenization. The extracted texts are denoted here by the symbol "| (Bengali full stop)". This segmentation divides the sentence, which is then trimmed by removing the extraneous white space. Python tokenization is done using the NLTK library method [28], [29].

2.4.2. Removing stop words

A word that does not contain any meaningful information in a sentence is called a stop word [9], [30]. In Bengali, these types of stop words are *এবং* (and), *ইহা* (it), *অনেক* (many), *তারপর* (then). As we mentioned earlier, a stop word dataset is created by merging and eliminating duplicate stop words from a variety of sources [31]. We have included 363 Bengali stop words in the dataset. Each sentence in the corpus was stored after the stop words were omitted from the sentence.

2.5. Generating similarity

In this phase, we have identified all the plagiarized sentences from the suspected documents as well as their counterparts in the source document. To detect plagiarism in a suspicious document, we have three steps: vectorization, similarity measure, and decision making. Figure 5 illustrates the cosine similarity measure procedure.

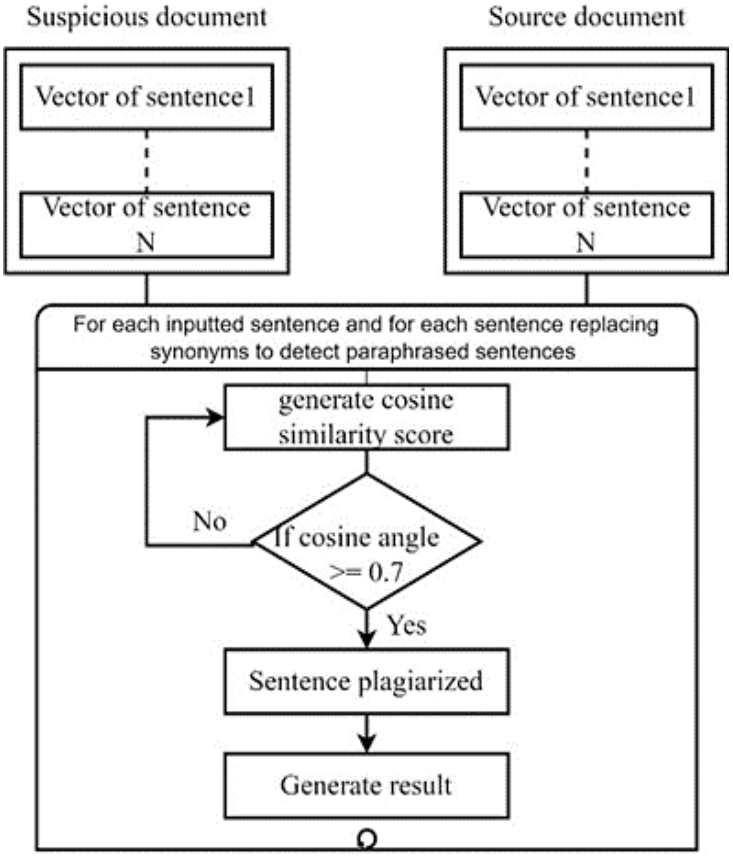

Figure 5. Cosine similarity measure

2.5.1. Vectorization

The TF-IDF method creates vector representations of sentences from the corpus. The suspect text must be processed in the same way as the corpus as shown in Figure 5. Firstly, the total number of tokens (unique words) in both manuscripts is calculated. To make the processing faster and more efficient, the TF-IDF results of each sentence are saved with the corpus sentence. Initially, *sklearn.feature_extraction* tool in Python was employed. However, this function could not effectively tokenize complicated Bengali words such as *প্রারম্ভ* (beginning), *কুরুক্ষেত্র* (battlefield), *কিংকর্তব্যবিমূঢ়* (bewildered), *লবডঙ্কা* (zero), and *লোকহিতৈষী* (altruistic). Therefore, we have implemented the TF-IDF algorithm from the sketch to improve the performance of the method.

2.5.2. Similarity measure

In this step, the TF-IDF scores of both the suspected and the source text are calculated using the cosine similarity method. As mentioned earlier, the cosine similarity algorithm gives the most accurate results over other algorithms on multibyte Unicode Bengali text. However, the Jaccard similarity algorithm [32] is also the closest performing algorithm for Bengali texts. Table 3 illustrates the proportion of similarity between articles 1 and 2.

- Article 1: *বাতাসে প্লাস্টিক, খাবারে প্লাস্টিক, এমনকি প্লাস্টিক পানিতেও। আর সেই প্লাস্টিক কণাই প্রতিদিন মানুষের দেহের ভেতরে ঢুকে বিষিয়ে দিচ্ছে শরীর। এতদিন পরিবেশের উপরে প্লাস্টিক দূষণের প্রভাব নিয়ে চিন্তায় ছিলেন বিজ্ঞানীরা।* (Plastic is in the air, plastic is in food, and plastic is in water as well. And these plastic particles are poisoning the human body every day. Until now, scientists have been concerned about the effects of plastic pollution on the environment).
- Article 2: *খাবারে প্লাস্টিক, বাতাসে প্লাস্টিক, এমনকি প্লাস্টিক পানিতেও। আর এই প্লাস্টিক প্রতিদিন মানুষের দেহের ভেতরে ঢুকে বিষিয়ে দিচ্ছে শরীর। বিজ্ঞানীরা এতদিন পরিবেশের উপরে প্লাস্টিক দূষণের প্রভাব নিয়ে চিন্তায় ছিলেন।* (Plastic is in the air, plastic is in food, and plastic is in water as well. And this plastic enters the human body every day and poisons the body. Scientists have long been concerned about the effects of plastic pollution on the environment).

Table 3. Compare the score of the algorithms

Cosine similarity	Jaccard similarity
0.83503	0.75815

The cosine similarity algorithm gives more precise results than the Jacquard similarity. In this case, if the statement is paraphrased by textual alterations or word or phrase correction, Jaccard's similarity does not improve the findings. As mentioned earlier, cosine similarity is a technique for calculating cosine angles between two non-zero vectors. Here, the results are limited to 0 and 1.

2.6. Decision making

A sentence is considered plagiarized when the similarity score between the suspect and the corpus vectors exceeds a certain threshold [33]. This study's threshold is 0.7% or 70%. We spent an ample amount of time testing different samples of Bengali articles to achieve this threshold. To generate a final plagiarism report, the algorithm collects plagiarized sentences, source names, similarity scores, and other necessary details. Each source's percentage of matches is stored and the copied sentences are also highlighted in red. If a source or collection of sources results in a very low level of match, the model calculates the total score of that match and displays the results under the label "*অন্যান্য* (others)". Finally, the system generates the final report of similarity with necessary information and references.

3. RESULTS AND DISCUSSION

This section discusses the experimental results and the analysis of the performance. Kaggle Notebook has been used as the experimental and computational environment in this project. This service provides 5 GB storage, 13 GB RAM, and 16 GB memory for Tesla P100 GPUs [34]. This model is tested with several suspicious documents. The proposed plagiarism detection system's functionality and output are demonstrated with two suspicious Bengali texts. The suspicious text-1 has been created by combining text from three books and a newspaper and therefore the entire paragraph is plagiarized.

Input: suspicious text 1

টেলিফোনের উদ্ভাবক আলেকজান্ডার গ্রাহাম বেল প্রথমে টেলিফোন সম্ভাষণ হিসেবে ''আহোয়'' শব্দটি ব্যবহার করেছিলেন, যা সাধারণত নাবিকেরা সম্ভাষণ হিসেবে ব্যবহার করতেন। পরে হ্যালো চালু হয়ে যায়। তার মানে আমি যে টেলিফোন ধরে কথা বলতে প্রস্তুত, সেটা জানিয়ে দেওয়া। একই সঙ্গে আমি কে, সেটাও জানানোর প্রথা রয়েছে। যেমন, ''হ্যালো? আমি অমুক বলছি'' / আমাদের ডাক ও টেলিযোগাযোগমন্ত্রীও প্রাথমিক স্তর থেকে কম্পিউটার প্রোগ্রামিং শিক্ষা শুরু করার কথা বলেছেন। রাশেদ ফারাজীর বাসনা এক- আধটা স্বপ্ন দেখা। নতুন কিছু দৃশ্য রচিত হোক। কিন্তু আক্ষেপ, ঘুমিয়ে সে কিছুই দেখে না। রাতের রাস্তায় দূরপাল্লার বাস চলছে। নানা ভঙ্গিতে ঘুমিয়ে থাকা যাত্রীরা দুলছে। বোয়িংয়ে কলকাতা থেকে ঢাকা ২৫ মিনিটের পথ।
(Alexander Graham Bell, the inventor of the telephone, first used the word "ahoy" as a telephone greeting, commonly used by sailors. After that, Hello is turned on. That means letting me know that I am ready to talk on the phone. At the same time, there is a custom to tell who I am. Like, "Hello? I say so and so". Our Posts and Telecommunication Minister has also said to start computer programming education from the elementary level. Rashed Farazi's wish is half a dream. Let some new scenes be composed. But alas, he sees nothing while sleeping. Long-distance buses are running on the night road. The sleeping passengers are swaying in various positions. 25 minutes flight from Kolkata to Dhaka by Boeing).

In Table 4, for the suspicious text-1, this system finds plagiarism in three major sources, "*তথ্য ও যোগাযোগ প্রযুক্তি - সপ্তম শ্রেণি* (Information and Communication Technology - Class VII)", "*কর্ম ও জীবনমুখী শিক্ষা - সপ্তম শ্রেণি* (Career and Life Oriented Education - Class VII)", and "*সাহিত্য কণিকা - অষ্টম শ্রেণি* (Literature - Class VIII)" as expected. Moreover, the system detected one minor source "*প্রথম আলো – সংবাদপত্র* (Prothom Alo – newspaper)" and some negligible sources combinedly named "*অন্যান্য* (other)". The suspicious text-2 has been compiled from a paragraph where half of the text was taken from two NCTB textbooks and the other half was taken from an original story written by an author of this paper.

Table 4. Final report of suspicious text article 1

Source of Article	Similarity Result
তথ্য ও যোগাযোগ প্রযুক্তি - সপ্তম শ্রেণি (Information and Communication Technology - Class VII)	38.46%
কর্ম ও জীবনমুখী শিক্ষা - সপ্তম শ্রেণি (Career and Life Oriented Education - Class VII)	15.38%
সাহিত্য কণিকা - অষ্টম শ্রেণি (Literature - Class VIII)	38.46%
প্রথম আলো – সংবাদপত্র (Prothom Alo – newspaper)	6.04%
অন্যান্য (Others)	1.66%
Total Similarity	100%

Input: suspicious text 2

অনেকেই ভাবে, যার মস্তিষ্ক যত বড়, তার চিন্তা করার শক্তি তত বেশি। ধারণাটি কিন্তু ভুল। মানুষের থেকে তিমির মস্তিষ্ক অনেক বড়। কিন্তু তিমির থেকে মানুষ বেশি বুদ্ধিমান। নেতৃত্ব, জ্ঞান, বিজ্ঞান, সাহিত্যচর্চা এসব বিষয়ে মানুষ উন্নতি করতে পেরেছে। বাংলা ব্যঞ্জনসন্ধির সূত্র মুখস্থ করতে করতে অনেকে অস্থির হয়ে যায়। সমীভবন হলো এমন এক প্রক্রিয়া, যেখানে এক ধ্বনির প্রভাবে পাশের ধ্বনিটি বদলে যায়। আমার ছোট ফুফুর বাড়ি। বহুদিন পর ফুফুর বাড়ি বেড়াতে এলাম। ২৪০ পাতার বইটিকে নিছক ভুয়ো বলে মানতে রাজি নন বিশেষজ্ঞরা। উর্দু ডানদিকে লেখা হয়েছে বইটি। ভেষজ বা আয়ুর্বেদিক চিকিৎসাবিদ্যার নানা দিক নিয়ে বিস্তৃত আলোচনা রয়েছে বইয়ে যা সত্যিই অবাক করে। সম্ভবত ১৯১২ সালে বইটি কিনে নেন বইয়ের ব্যবসা করা উইলফ্রিড ভয়নিচ। তার নামেই বইটির এই নামকরণ হয়েছে। এই মতিভ্রম ওর কেন হলো, কিছুই বুঝতে পারছি না। দ্বীপ জুড়ে নেই কোনো গাছ বা লতা-গুল্ম। আগামী মাসে প্রথম বর্ষের পরীক্ষায় অংশ নেওয়ার কথা ছিল তাঁর। তাপসকে খুব কাছ থেকে দেখেছেন যোগাযোগ ও সাংবাদিকতা বিভাগের শিক্ষার্থী পার্থ বণিক।
(Many people think that the bigger the brain, the greater the thinking power. But the idea is wrong. Whale brains are much bigger than humans. But humans are smarter than whales. People have been able to improve in leadership, knowledge, science, and literary practice. Many people get restless while memorizing the formula of Bengali consonants. Assimilation is a process in which one sound changes the sound next to it. My younger cousin's house. After many days, I came to visit my uncle's house. Experts are not willing to accept the 240-page book as a mere fake. The book is written in Urdu on the right side. The book contains extensive discussions on various aspects of herbal or Ayurvedic medicine which is really surprising. The book was probably bought by Wilfried Voynich, a bookseller, in 1912. The book is named after him. I do not understand why this delusion happened to him. There are no trees or vines on the island. He was supposed to appear in the first-year examination next month. Partha Vanik, a student of the Communication and Journalism Department, saw Tapas very closely).

Two significant sources of similarity are depicted in Table 5 i.e., "*বাংলা সাহিত্য - নবম ও দশম শ্রেণি* (Bengali Literature - 9th and 10th Class)", "*বাংলা ব্যাকরণ ও নির্মিতি - নবম ও দশম শ্রেণি* (Bengali Grammar and Construction - 9th and 10th Class)", and a minor combined source "*অন্যান্য* (others)". This indicates that the passage has a similarity score of 48.81 percent as expected. Figure 6 illustrates the experimental result.

Table 5. Final report of suspicious text article 2

Source of Article	Similarity Result
বাংলা সাহিত্য - নবম ও দশম শ্রেণি (Bengali Literature - 9th and 10th Class)	36.47%
বাংলা ব্যাকরণ ও নির্মিতি - নবম ও দশম শ্রেণি (Bengali Grammar and Construction - 9th and 10th Class)	11.64%
অন্যান্য (Others)	0.7%
Total Similarity	48.81%

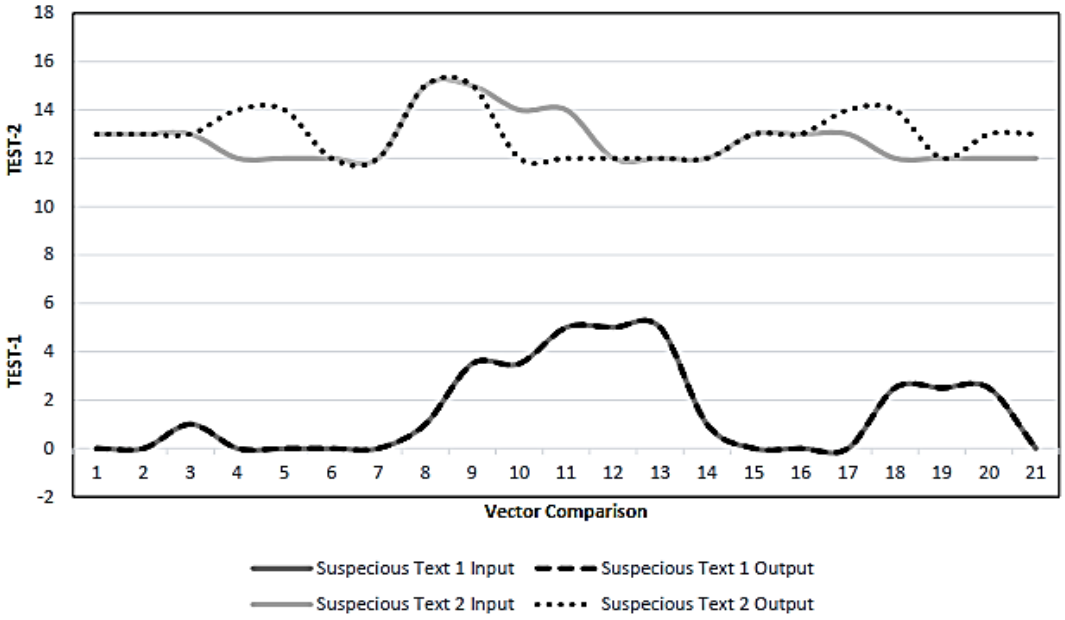

Figure 6. Experimental result analysis

A test dataset of 400 Bengali passages has been used to verify the method. The system generates a report with 386 correct cases and 14 minor erroneous reports and that concludes the accuracy rate of the proposed method is 97.31%. The reason behind erroneous results is the scarcity of data. Although the proposed method is extrinsic and it has covered only two domains (i.e., education and newspaper), it lacks a large scale of data in the corpus. This leads to some erroneous results on our test Bengali articles.

4. CONCLUSION

This paper demonstrates an extrinsic monolingual plagiarism detection approach for Bengali texts. The proposed method generates vectors of each sentence using the TF-IDF vectorizer that is later used to find fully or partially similar sentences in the source and suspicious text. The work is limited to educational and newspaper domains at this phase. Although our primary goal has been accomplished successfully and achieved a satisfactory accuracy of 97.31%, it has some limitations as well. To overcome the limitations, the proposed method will be improved in the future by distinguishing between passive and active sentences, expanding the semantic knowledge base, and determining semantic and word-order similarity between two sentences. Authors might also work on detecting plagiarism without a reference in the future.

ACKNOWLEDGEMENTS

The project is funded by Institute for Advanced Research (IAR) Publication Grant of United International University (UIU), Bangladesh (Ref. No.- IAR/2022/Pub/017). The authors are very grateful to IAR, UIU.

REFERENCES

[1] M. AlSallal, R. Iqbal, V. Palade, S. Amin, and V. Chang, "An integrated approach for intrinsic plagiarism detection," *Future Generation Computer Systems*, vol. 96, pp. 700–712, Jul. 2019, doi: 10.1016/j.future.2017.11.023.

[2] A. Abdi, S. M. Shamsuddin, N. Idris, R. M. Alguliyev, and R. M. Aliguliyev, "A linguistic treatment for automatic external plagiarism detection," *Knowledge-Based Systems*, vol. 135, pp. 135–146, Nov. 2017, doi: 10.1016/j.knosys.2017.08.008.

[3] J. P. Wahle, T. Ruas, T. Foltýnek, N. Meuschke, and B. Gipp, "Identifying machine-paraphrased plagiarism," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13192, Springer International Publishing, 2022, pp. 393–413.

[4] V. Pupovac, "The frequency of plagiarism identified by text-matching software in scientific articles: a systematic review and meta-analysis," *Scientometrics*, vol. 126, no. 11, pp. 8981–9003, Nov. 2021, doi: 10.1007/s11192-021-04140-5.

[5] M. Najm Mansoor and M. S. H. Al-Tamimi, "Computer-based plagiarism detection techniques: A comparative study," *International Journal of Non-linear Analysis and Applications (IJNAA)*, vol. 13, pp. 2008–6822, 2022.

[6] N. Islam, M. M. Hoque, and M. R. Hossain, "Automatic authorship detection from Bengali text using stylometric approach," in *20th International Conference of Computer and Information Technology, ICCIT 2017*, Dec. 2018, vol. 2018-January, pp. 1–6, doi: 10.1109/ICCITECHN.2017.8281793.

[7] R. Pandit, S. Sengupta, S. K. Naskar, N. S. Dash, and M. M. Sardar, "Improving semantic similarity with cross-lingual resources: a study in Bangla—a low resourced language," *Informatics*, vol. 6, no. 2, May 2019, doi: 10.3390/informatics6020019.

[8] M. F. Mridha, A. Q. Ohi, J. Shin, M. M. Kabir, M. M. Monowar, and M. A. Hamid, "A Thresholded Gabor-CNN Based Writer Identification System for Indic Scripts," *IEEE Access*, vol. 9, pp. 132329–132341, 2021, doi: 10.1109/ACCESS.2021.3114799.

[9] N. Hossain, H. R. Milon, S. N. U. Sabbir, and A. Inan, "Inclusive bidirectional conversion system between Chittagonian and standard Bangla," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 11, no. 1, pp. 396–404, Feb. 2022, doi: 10.11591/eei.v11i1.3237.

[10] A. Oktoveri, A. T. Wibowo, and A. M. Barmawi, "Non-relevant document reduction in anti-plagiarism using asymmetric similarity and AVL tree index," in *2014 5th International Conference on Intelligent and Advanced Systems (ICIAS)*, Jun. 2014, pp. 1–5, doi: 10.1109/ICIAS.2014.6869547.

[11] P. Mahdavi, Z. Siadati, and F. Yaghmaee, "Automatic external Persian plagiarism detection using vector space model," in *Proceedings of the 4th International Conference on Computer and Knowledge Engineering, ICCKE 2014*, Oct. 2014, pp. 697–702, doi: 10.1109/ICCKE.2014.6993398.

[12] N. Riya Ravi, K. Vani, and D. Gupta, "Exploration of fuzzy C means clustering algorithm in external plagiarism detection system," in *Advances in Intelligent Systems and Computing*, vol. 384, Springer International Publishing, 2016, pp. 127–138.

[13] M. Paul and S. Jamal, "An improved SRL based plagiarism detection technique using Sentence ranking," *Procedia Computer Science*, vol. 46, pp. 223–230, 2015, doi: 10.1016/j.procs.2015.02.015.

[14] A. Rahmatulloh, N. I. Kurniati, I. Darmawan, A. Z. Asyikin, and D. Witarsyah J, "Comparison between the stemmer porter effect and Nazief-Adriani on the performance of winnowing algorithms for measuring plagiarism," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 9, no. 4, pp. 1124–1128, Aug. 2019, doi: 10.18517/ijaseit.9.4.8844.

[15] M. Roostaee, S. M. Fakhrahmad, and M. H. Sadreddini, "Cross-language text alignment: A proposed two-level matching scheme for plagiarism detection," *Expert Systems with Applications*, vol. 160, Dec. 2020, doi: 10.1016/j.eswa.2020.113718.

[16] E. M. Hambi and F. Benabbou, "A deep learning based technique for plagiarism detection: A comparative study," *IAES International Journal of Artificial Intelligence*, vol. 9, no. 1, pp. 81–90, Mar. 2020, doi: 10.11591/ijai.v9.i1.pp81-90.

[17] W. Darmalaksana, C. Slamet, W. B. Zulfikar, I. F. Fadillah, D. S. Maylawati, and H. Ali, "Latent semantic analysis and cosine similarity for hadith search engine," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 18, no. 1, Feb. 2020, doi: 10.12928/telkomnika.v18i1.14874.

[18] M. Kaur, V. Gupta, and R. Kaur, "Semantic-based integrated plagiarism detection approach for English documents," *IETE Journal of Research*, pp. 1–17, Dec. 2021, doi: 10.1080/03772063.2021.2004383.

[19] F. Alvi, M. Stevenson, and P. Clough, "Paraphrase type identification for plagiarism detection using contexts and word embeddings," *International Journal of Educational Technology in Higher Education*, vol. 18, no. 1, Aug. 2021, doi: 10.1186/s41239-021-00277-8.

[20] N. Van Son, L. T. Huong, and N. C. Thanh, "A two-phase plagiarism detection system based on multi-layer LSTM networks," *IAES International Journal of Artificial Intelligence*, vol. 10, no. 3, pp. 636–648, Sep. 2021, doi: 10.11591/ijai.v10.i3.pp636-648.

[21] O. Hourrane and E. H. Benlahmar, "Graph transformer for cross-lingual plagiarism detection," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 3, pp. 905–915, Sep. 2022, doi: 10.11591/ijai.v11.i3.pp905-915.

[22] M. Sabeeh and F. Khaled, "Plagiarism detection methods and tools: An overview," *Iraqi Journal of Science*, pp. 2771–2783, Aug. 2021, doi: 10.24996/ijs.2021.62.8.30.

[23] H. Veisi, M. Golchinpour, M. Salehi, and E. Gharavi, "Multi-level text document similarity estimation and its application for plagiarism detection," *Iran Journal of Computer Science*, vol. 5, no. 2, pp. 143–155, Jun. 2022, doi: 10.1007/s42044-022-00098-6.

[24] A. Jalilifard, V. F. Caridá, A. F. Mansano, R. S. Cristo, and F. P. C. da Fonseca, "Semantic sensitive TF-IDF to determine word relevance in documents," in *Lecture Notes in Electrical Engineering*, vol. 736, Springer Singapore, 2021, pp. 327–337.

[25] "National Curriculum and Textbook Board (NCTB)," (in Bengali), NCTB, http://www.nctb.gov.bd/ (accessed Jan. 11, 2022).

[26] N. Hossain, S. Islam, and M. N. Huda, "Development of Bangla spell and grammar checkers: resource creation and evaluation," *IEEE Access*, vol. 9, pp. 141079–141097, 2021, doi: 10.1109/ACCESS.2021.3119627.

[27] Ahmed Sharif, *Bangla academy samkshipta bangla abhidhan*, 2nd ed. Bangla Academy, 2020.

[28] M. Wang and F. Hu, "The application of NLTK library for Python natural language processing in Corpus research," *Theory and Practice in Language Studies*, vol. 11, no. 9, pp. 1041–1049, Sep. 2021, doi: 10.17507/tpls.1109.09.

[29] T. B. Shahi and C. Sitaula, "Natural language processing for Nepali text: a review," *Artificial Intelligence Review*, vol. 55, no. 4, pp. 3401–3429, Apr. 2022, doi: 10.1007/s10462-021-10093-1.

[30] J. Leskovec, A. Rajaraman, and J. D. Ullman, *Mining of massive datasets*. Cambridge University Press, 2020.

[31] "Bengali Stopwords-Ranks NL webmaster tools," Ranks NL. https://tinyurl.com/stopword-bn-ranks (accessed Jan. 11, 2022).

[32] T. Wahyuningsih, "Text mining an automatic short answer grading (ASAG), comparison of three methods of cosine similarity, Jaccard similarity and Dice's coefficient," *Journal of Applied Data Sciences*, vol. 2, no. 2, May 2021, doi: 10.47738/jads.v2i2.31.

[33] J. Muangprathub, S. Kajornkasirat, and A. Wanichsombat, "Document plagiarism detection using a new concept similarity in formal concept analysis," *Journal of Applied Mathematics*, vol. 2021, pp. 1–10, Mar. 2021, doi: 10.1155/2021/6662984.

[34] A. Y. Wang *et al.*, "What makes a well-documented notebook? a case study of data scientists' documentation practices in Kaggle," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, May 2021, pp. 1–7, doi: 10.1145/3411763.3451617.

BIOGRAPHIES OF AUTHORS

Adil Ahnaf graduated from United International University (UIU) with a bachelor's degree in computer science and engineering (CSE). His research interests are primarily in the area of natural language processing, machine learning, and deep learning. He is the author/co-author of 2 research publications. Currently working as a freelancer on Fiverr. He can be contacted at aahnaf151054@bscse.uiu.ac.bd.

Hossain Mohammad Mahmudul Hasan graduated from UIU with a Bachelor of Science in CSE. He is currently working as a full-stack developer at Brac It Services Ltd. He received an Hackathoon Champion Honor in 2017 in Bracathon's 2nd Installment. His research interests are primarily in the area of artificial intelligence, data analytics and natural language processing, where he is the author/co-author of 3 research publications. He can be contacted at hhasan152041@bscse.uiu.ac.bd.

Nabila Sabrin Sworna is a lecturer in the Department of CSE of United International University, Bangladesh. She has a bachelor's degree from the same background. Her research interests include machine learning, data mining, pattern recognition, and natural language processing. She can be contacted at nabila@cse.uiu.ac.bd.

Nahid Hossain is an assistant professor of the Computer Science and Engineering (CSE) Department of United International University (UIU), Bangladesh. Before joining UIU, he worked as a software engineer in the Natural Language Processing (NLP) Department at eGeneration Ltd, Bangladesh. He has a master's and a bachelor's degree in the same background. He has got several national and international scholarships and awards including a scholarship from the European Union and a Gold Medal from the Education Minister of Bangladesh. His research interests include mainly natural language processing and machine learning. He can be contacted at nahid@cse.uiu.ac.bd, and his profile can also be found at www.nahid.org.